# Controlling the Danger of False Discoveries in Estimating Multiple Treatment Effects

Dan Wunderli *
Department of Economics, University of Zurich
Wilfriedstrasse 6, CH-8006 Zurich, Switzerland
dan.wunderli@econ.uzh.ch

December 2011

**Abstract**

I expose the risk of false discoveries in the context of multiple treatment effects. A false discovery is a nonexistent effect that is falsely labeled as statistically significant by its individual $t$-value. Labeling nonexistent effects as statistically significant has wide-ranging academic and policy-related implications, like costly false conclusions from policy evaluations. I reexamine an empirical labor market model by using state-of-the art multiple testing methods and I provide simulation evidence. By merely using individual $t$-values at conventional significance levels, the risk of labeling probably nonexistent treatment effects as statistically significant is unacceptably high. Individual $t$-values even label a number of treatment effects as significant, whereas multiple testing indicates false discoveries in these cases. Tests of a joint null hypothesis such as the well-known $F$-test control the risk of false discoveries only to a limited extent and do not optimally allow for rejecting individual hypotheses. Multiple testing methods control the risk of false discoveries in general while allowing for individual decisions in the sense of rejecting individual hypotheses.

KEY WORDS: False discoveries, multiple error rates, multiple treatment effects, labor market

JEL CLASSIFICATION NOS: C12, C14, C21, C31, C41, J08, J64.

*Many thanks to Rafael Lalive, Ashok Kaul, Rainer Winkelmann, and Michael Wolf for their comments. Thanks also go to Gregori Baetschmann, Alexandru Popescu, and participants of the microeconometrics research seminar at the University of Zurich for useful comments.

# 1 Introduction

I put the danger of false discoveries into perspective by providing simulation evidence and by reexamining treatment effects within an empirical labor market model of Lalive et al. (2005). A false discovery is a nonexistent effect that is falsely labeled as statistically significant by its individual $t$-value. I provide evidence that the risk of making false discoveries is unacceptably high if one does not account for the danger of false discoveries. It is shown that individual $t$-values even label a number of treatment effects as statistically significant that are probably false discoveries. As usual in inferential statistics, one can only 'prove beyond a reasonable doubt' that an effect exists. One can show by multiple testing methods that some individually significant treatment effects are probably nonexistent, as quantified in a so-called multiple significance level. In this paper, 'nonexistent treatment effects' should be understood in this inferential way.

In the empirical labor market model of Lalive et al. (2005), there are multiple treatment effects. These treatment effects are potentially interrelated, thus one must consider the treatment effects jointly. The central empirical question is whether some of the treatment effects being individually significant are false discoveries in the sense of not having controlled the risk of labeling nonexistent treatment effects as statistically significant, that is, of making false discoveries. I reexamine the empirical model of Lalive et al. (2005) with respect to making false discoveries. I control the risk of labeling nonexistent treatment effects as (individually) significant by using the powerful multiple testing methods from Romano and Wolf (2005).

To ease understanding, I chose the somewhat vague phrase 'one must consider the effects *jointly*' instead of the technically correct 'one must consider effects with a *multiple* error type one'. In reading jointly, most readers probably think of an $F$-test. However, an $F$-test cannot control the risk of labeling some nonexistent treatment effects as statistically significant. To see why, let us first consider the difference between testing a number of individual hypotheses and testing one joint null hypothesis; the difference between the former and multiple testing is explained in a second step. The crucial first point is that a joint null hypothesis does not allow for individual decisions in the sense of rejecting individual null hypotheses from a joint point of view.

**Number of individual hypotheses versus one joint hypothesis** For the sake of exposition, consider the regression

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \epsilon_i. \tag{1}$$

Suppose that $\beta_1, \beta_2, \beta_3$ measure treatment effects. It is clear that testing one *joint* null hypothesis $H_{0,joint} : \beta_1 = \beta_2 = \beta_3 = 0$ with an $F$-test does not necessarily yield the same result than testing *three single* null hypotheses $H_{0,1} : \beta_1 = 0$, $H_{0,2} : \beta_2 = 0$, and $H_{0,3} : \beta_3 = 0$ in the following sense. It may be the case that the first and third individual hypothesis $H_{0,1}$, $H_{0,3}$ are rejected, although the joint $H_{0,joint}$ cannot be rejected. Nonetheless, some coefficients may still be significant by jointly considering the three coefficients, despite the fact that the $F$-test could not reject $H_{0,joint} : \beta_1 = \beta_2 = \beta_3 = 0$. Furthermore, it may happen that no individual $t$-test rejects while the joint $F$-test rejects, as for highly correlated regressors. By $t$-testing I mean calculating the $t$istic and comparing it to the appropriate quantile of the $t$ (or normal) distribution. Of course, simply relying on individual $t$-tests of single null hypotheses and bluntly discarding the joint point of view of the $F$-test is not the solution. The more general Wald, Lagrange Multiplier (LM), or Likelihood Ratio (LR) tests have the same drawback than

the $F$-test. They can just test one *joint* nonlinear hypothesis $H_{0,joint} : \mathbf{a}(\theta_0) = \mathbf{0}$ against its joint alternative hypothesis[1]. We really want a test that considers statistical significance jointly, as in the $F$-test. But we also want this joint test to tell us which individual coefficients out of the joint null hypothesis are individually significant while taking into account their joint nature. Section 3 provides graphical illustrations in this respect. We want a joint test in which individual rejections are possible.

**Individual testing versus multiple testing** To this end, multiple testing methods tell us exactly *which* null hypotheses out of the family of individual null hypotheses can be rejected at a given *multiple* significance level. A *multiple* significance level takes account of the danger of false discoveries. None of the aforementioned tests of the joint null hypothesis can optimally tell us *which single* coefficients are statistically significant as seen from a point of view joint with the other coefficients under scrutiny. Table 1 explains individual versus joint versus multiple testing for the case of $p > 1$ regression coefficients.

| | Risk of false discoveries controlled | Individual decision possible | Number of null hypotheses |
|---|---|---|---|
| Individual tests | No | Yes | $p > 1$[a] |
| $F$-test or Wald, LM, LR | Maybe | No[b] | one[c] |
| Multiple testing | Yes | Yes | $p > 1$[d] |

[a]There is one null hypothesis for each of the $p$ coefficients $H_{0,s} : \beta_s = 0$ for $s = 1, \ldots, p$.
[b]Rectangular approximations to the elliptic joint confidence region are possible. See Section 3.
[c]All individual coefficients are merged to one *joint* null hypothesis $H_{0,joint} : \beta_1 = \ldots = \beta_p = 0$.
[d]All individual coefficients are merged to a *family* of null hypotheses $\{H_{0,s} : \beta_s = 0,\ s = 1, \ldots, p\}$.

Table 1: Multiple Testing allows joint testing while individual decisions are possible

The crucial point is that only multiple testing methods can generally control the risk of false discoveries, while optimally allowing for rejecting individual hypotheses from a joint point of view.

In our stylized regression example (1), suppose as before that only $\beta_1$ and $\beta_3$ are individually significant according to $t$-testing $\beta_1$, $\beta_2$, and $\beta_3$ at the 5% level. Regardless of the outcome of testing the joint $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ at the 5% level, it may well be the case that a multiple error type one is very large instead of being at a conventional level between 1% and 10%. In this paper, I consider a multiple error type one called the familywise error rate FWE that is defined as follows

$$\begin{aligned} \text{Familywise error rate FWE} &= P(\text{Number of falsely rejected null hypotheses} \geq 1) \\ &= P(\text{Number of false discoveries} \geq 1). \end{aligned} \tag{2}$$

where $P(\cdot)$ denotes the probabilty mechanism. The well-known error type one '$\alpha$' of an individual test is the probability of falsely rejecting the null hypothesis $H_0 : \beta_s = 0$ for one specific $s \in \{1, \ldots, p\}$. Thus, it makes sense that the multiple error type one FWE is an error probability as well. Controlling the error probability FWE at the 5% level means ensuring that FWE $\leq$ 5%. If the familywise error rate is not explicitly taken care of, it may easily be the case that only FWE $\leq$ 50% holds. Section 2 provides a Monte Carlo simulation in this respect.

Note that a familywise error rate FWE at 50% instead of at 5% renders statistical inference useless. One cannot trust in the comparison of $t$-values to the conventional critical value $c_{1-\alpha/2}^{N(0,1)}$, or to its

[1]Where $\mathbf{a} : \mathbb{R}^p \to \mathbb{R}^p$ is a continuously differentiable function

bootstrap version $c^{boot}_{1-\alpha/2}$, in presence of such a high probability of labeling one or more nonexistent effects as statistically significant. In the empirical results of this paper, we will even see an empirical instance where carrying out six individual $t$-tests at the 5% level translates to the FWE being around 90%, that is, the probability of having made one or more false discoveries is around 90%.

**Why False Discoveries Matter** A high familywise error rate FWE translates into a high risk of having labeled some treatment effects as individually significant that do not exist, which statisticians call false discoveries. The larger the family of individual null hypotheses is that are scrutinized jointly, the higher is the risk of labeling some treatment effects as significant that do not exist. Thus, one runs the uncontrolled danger of so-called false discoveries by testing merely individually.

Not controlling for false discoveries has wide-ranging academic and policy-related implications. In policy evaluation, if effects are falsely labeled as significant, wrong policies may be pursued, leading to a waste of public funds or to an unexpected deterioration where an improvement was expected. False discoveries are sometimes even published results as if there had been no prescreening of results based on individual $p$-values. Heckman et al. (2010) refer to this problem as 'cherry picking'. In this sense, results that 'did not work' should be reported along with results that 'worked'. The common robustness checks that report results of different specifications certainly are steps in the right direction.

The remainder of this paper is organized as follows. Section 2 provides simulation evidence on how large the danger of false discoveries can be by testing merely individually. Section 3 explains the difference between testing one joint null hypothesis and multiple testing of a family of hypotheses with some graphs. Section 4 briefly summarizes the labor market model in Lalive et al. (2005) that I reexamine. A more detailed description of the model is in Appendix A. Section 5 describes different extents of detail in distinguishing treatment effects and reports results from individual and multiple testing of these treatment effects. Section 5 thus indicates to which extent the danger of false discoveries is ignored by individual $t$-tests within the empirical labor market model of Lalive et al. (2005). Section 6 concludes.

# 2 Individually Testing $p > 1$ Null Hypotheses versus Multiple Testing

## 2.1 Simulation Setup

The point that I illustrate in this section is: How large can the probability of falsely rejecting one or more null hypotheses (FWE) be by naively testing all $p$ null hypotheses merely individually?

Consider the following simulation setup. There are $p$ explanatory random variables $X_1, \ldots, X_p$, with which one associates $p$ treatment effects $\beta_1, \ldots, \beta_p$ onto the random response variable $Y$. The explanatory variables may be correlated with each other. The data generating process is

$$Y = c + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon_t, \tag{3}$$

where $\epsilon_t \sim N(0,1)$. There is not a single (nonzero) treatment effect: $\beta_1 = \ldots = \beta_p = 0$.

The empiricist only observes data sets of size $N$ from the data generating process (3), resulting in estimates $\hat{\beta}_1, \ldots, \hat{\beta}_p$ that may assume nonzero values based on an observed sample. Given that there does not exist any treatment effect, a test that labels any $\hat{\beta}_s$, $s = 1, \ldots, p$, as statistically significant commits an error type one; any sensible test at level $\alpha$ ensures that this error does not occur more

often than $\alpha \cdot M$ times, at least asymptotically as the number of data sets $M \to \infty$. Let us check if the conventional $t$-test at level $\alpha$ falsely labels nonexistent treatment effects as significant in no more than $\alpha \cdot M$ cases, for the null hypothesis $H_{0,s} : \beta_s = 0$ versus alternative $H_{A,s} : \beta_s \neq 0$ for each $s = 1, \ldots, p$.

First, I generate $M = 2000$ data sets of size $N$ from the data generating process (3), denoted as $x^{(1)}, \ldots, x^{(M)}$. Second, I compute the estimates $\hat{\beta}_1^{(m)}, \ldots, \hat{\beta}_p^{(m)}$ based on each generated data set $x^{(m)}$, as well as the respective $t$-statistics $t_1^{(m)}, \ldots, t_p^{(m)}$. $H_{0,s}$ is rejected on data set $x^{(m)}$ if $|t_s^{(m)}| > c_{1-\alpha/2}$ holds. If one or more $t$-tests reject on data set $x^{(m)}$, which is the case if the number of rejections $\sum_{s=1}^{p} \mathbf{1}[|t_s^{(m)}| > c_{1-\alpha/2}]$ is larger than one, a familywise error $FErr^{(m)}$ is committed on data set $x^{(m)}$

$$FErr^{(m)} = \mathbf{1}\left[\sum_{s=1}^{p} \mathbf{1}[|t_s^{(m)}| > c_{1-\alpha/2}] > 1\right] \tag{4}$$

The estimated probability of falsely labeling one or more $\hat{\beta}_s$ as statistically significant (committing a $FErr$) at level $\alpha$ is the arithmetic mean over all $M$ simulation runs

$$FWE = \frac{1}{M} \sum_{m=1}^{M} FErr^{(m)} \tag{5}$$

Clearly, FWE $\approx \alpha$ should hold if individually $t$-testing $p$ treatment effects should be any help in controlling the danger of labeling nonexistent treatment effects as statistically significant. I check this in a number of cases. Namely, let the number of (nonexistent) treatment effects $p$ lie in $\{2, 5, 10, 20\}$. I allow the explanatory variables $[X_1, \ldots, X_p]'$ to be correlated with each other, according to the covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{p-1} \\ \rho & 1 & \rho & \ldots & \rho^{p-2} \\ \rho^2 & \rho & 1 & \ldots & \rho^{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & \ldots & 1 \end{pmatrix}, \quad \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \sim N(\mathbf{0}, \mathbf{\Sigma}) \tag{6}$$

for $\rho \in \{0.9, 0.5, 0, -0.5, -0.9\}$. That is, the further the indices $s, q$ of two explanatory variables $X_s, X_q$ are apart, the less $X_s, X_q$ are correlated with each other in absolute value. A data set of explanatory variables $x^{(m)}$ can then be simulated from a set of $p$ i.i.d. $N(0, 1)$ variables by using the triangular Cholesky factor $\mathbf{C}$ as in $\mathbf{\Sigma} = \mathbf{C}\mathbf{C}'$.

## 2.2 Simulation Results

Table 2 summarizes the results from simulation setup 2.1 for: numbers of (nonexistent) treatment effects $p \in \{2, 5, 10, 20\}$, correlation $\rho \in \{0.9, 0.5, 0, -0.5, -0.9\}$, size of data sets $N = 1000$ [2], level of individual tests $\alpha = 5\%$, and number of simulation runs $M = 2000$.

First, note that the higher the number of treatment effects $p$, the higher is the probability of discovering one or more (nonexistent) treatment effects. Hence, the higher the number of multiple treatment effects is, the higher is the danger of making one or more false discoveries. Note that even for only two treatment effects $p = 2$, the FWE is 7.1% at best instead of $\alpha = 5\%$, which a sensible (multiple) test ought to ensure. For $p = 5$ and larger, the probability of falsely labeling one or more nonexistent effects as significant rises to 63.2%.

[2] For $N = 100$, the results are virtually identical

| FWE | $\rho = 0.9$ | $\rho = 0.5$ | $\rho = 0$ | $\rho = -0.5$ | $\rho = -0.9$ |
|---|---|---|---|---|---|
| $p = 2$ | 7.3% | 9.6% | 10.1% | 9.1% | 7.1% |
| $p = 5$ | 19.4% | 21.4% | 23.2% | 21.6% | 19.8% |
| $p = 10$ | 36.4% | 38.9% | 40.3% | 38.3% | 36.7% |
| $p = 20$ | 57.8% | 60.2% | 63.2% | 60.1% | 57.9% |

Table 2: Probability of falsely labeling one or more nonexistent treatment effects out of $p$ as statistically significant by doing $p$ $t$-tests individually at level $\alpha = 5\%$

Second, observe that the probability of labeling nonexistent treatment effects as significant (FWE) is highest for uncorrelated treatment effects $\rho = 0$. The higher the correlation between the $p$ explanatory variables is in absolute value, the lower is the probability of committing a familywise error, which is not surprising. In case of perfect correlation $\rho = 1$, one essentially tests only one null hypothesis in individually testing all $p$ null hypotheses $\{H_{0,s} : \beta_s = 0,\ s = 1, \ldots, p\}$; knowing one $\hat{\beta}_s$ means knowing all other $\hat{\beta}_q$, $s \neq q$. For uncorrelated $X_1, \ldots, X_p$, however, $\hat{\beta}_s$ is unrelated to $\hat{\beta}_q = 0$ for $s \neq q$: thus each single $\hat{\beta}_s$, $s = 1, \ldots, p$, poses a danger of making a false discovery.

Third, there does not seem to be a systematic pattern with respect to positive or negative correlation $\rho$, the latter meaning alternating between negative and positive correlation as in row $[1, \rho, \rho^2, \ldots, \rho^{p-1}]$ of $\mathbf{\Sigma}$ [3].

Bear in mind that the probability of labeling one or more nonexistent treatment effects as significant can be much higher for a given number of treatment effects $p$ than in this simple simulation study. The empirical part provides an example where the FWE is 90% for $p = 6$ treatment effects.

This section illustrated that individual confidence intervals at level $\alpha$ do not control the probability of labeling one or more treatment effects at level $\alpha$. Hence, the confidence intervals resulting from naively testing $s = 1, \ldots, p$ individual null hypotheses under $N(0, 1)$ or under bootstrapping

$$\text{naive } CI_{\beta_s}^{N(0,1)} = [\hat{\beta}_s - \hat{\sigma}_{\beta_s} c_{1-\alpha}^{N(0,1)},\ \hat{\beta}_s + \hat{\sigma}_{\beta_s} c_{1-\alpha}^{N(0,1)}] \tag{7}$$

$$\text{naive } CI_{\beta_s}^{boot} = [\hat{\beta}_s - \hat{\sigma}_{\beta_s} c_{1-\alpha}^{boot},\ \hat{\beta}_s + \hat{\sigma}_{\beta_s} c_{1-\alpha}^{boot}] \tag{8}$$

are too small. In geometric terms, this being too small means that the rectangular region spanned by all $s = 1, \ldots, p$ individual naive $CI_{\beta_s}^{N(0,1)}$ or naive $CI_{\beta_s}^{boot}$ has coverage probability smaller than $1 - \alpha$.

## 3 Tests of One Joint Null Hypothesis versus Multiple Testing

One of the key issues in comparing tests of one joint null hypothesis to multiple testing is whether individual decisions in the sense of rejecting individual null hypotheses are possible, as summarized in Table 1. From a purely mechanical point of view, individual decisions are possible with an $F$-test, say, by projecting the ellipsoidal confidence region onto the axes. However, the $H_{0,joint}$-derived individual confidence intervals are too large. In multiple testing, however, one sets up a rectangular joint confidence region, leading to small individual confidence intervals with favorable size and power properties.

[3] For all off-diagonal elements in $\mathbf{\Sigma}$ being negative, the Cholesky factor does not exist because $\mathbf{\Sigma}$ is not positive semi-definite in this case.

To see why, consider the case of a joint confidence region for the two-dimensional parameter $(\beta_1, \beta_2)$ from a linear regression such as in equation (1), representing two treatment effects. In multiple testing, one constructs a rectangular joint confidence region, as depicted in Figure 1. Thus, individual decisions based on this rectangular joint confidence region are straight forward. One just projects each side of the rectangle onto the corresponding axis, where small confidence intervals result. Specifically, one finds one single multiple critical value $c_{1-\alpha}^{MTest}$ in multiple testing. The individual confidence intervals directly inferred from the rectangular joint confidence region are

$$\text{individual } CI_{\beta_s}^{MTest} = [\hat{\beta}_s - \hat{\sigma}_{\beta_s} c_{1-\alpha}^{MTest},\ \hat{\beta}_s + \hat{\sigma}_{\beta_s} c_{1-\alpha}^{MTest}] \tag{9}$$

for each individual parameter $\beta_s$. In the $\beta_1, \beta_2$ example above, the resulting $CI_{\beta_1}^{MTest}$ and $CI_{\beta_2}^{MTest}$ are large enough, such that the familywise error rate FWE as in (2) is controlled. But $CI_{\beta_1}^{MTest}$ and $CI_{\beta_2}^{MTest}$ are also small enough so that a rejection occurs with a high probability when $H_{0,s} : \beta_s = 0$ is wrong, meaning that the test has high statistical power.

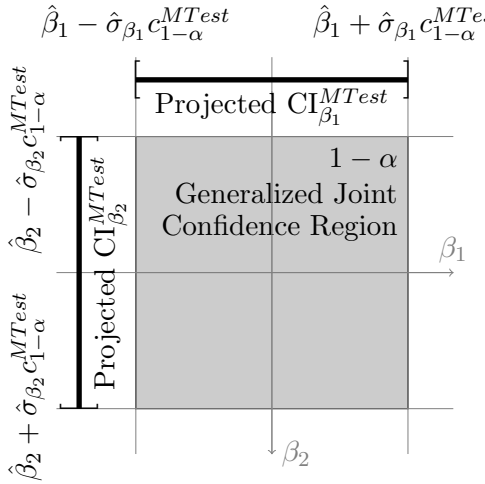


Figure 1: Individual CI's inferred from rectangular joint confidence region as in Romano and Wolf (2005)'s multiple testing method

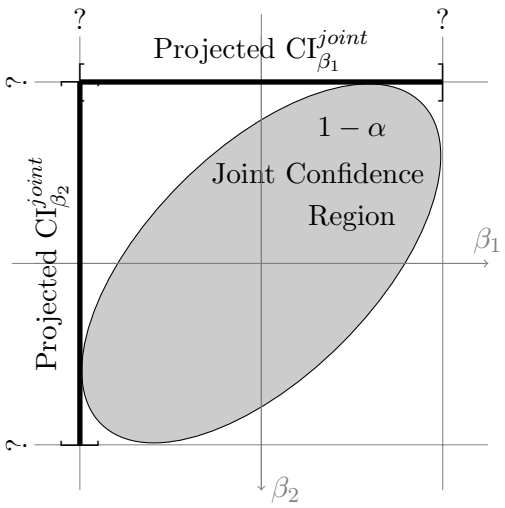


Figure 2: Individual CI's inferred from ellipsoidal joint confidence region as in a test of one joint null hypothesis, e.g. $F$-test.

In setting up a joint null hypothesis $H_{0,joint} : \beta_1 = \beta_2 = 0$ and $F$-testing it at significance level $\alpha$, we implicitly set up an ellipsoidal joint confidence region as depicted in Figure 2. The more general Wald, LM, LR tests make use of ellipsoidal confidence regions as well, by considering an ellipsoidal confidence region at level $1-\alpha$ whose ellipsoidal boundary is characterized by points with a constant $\chi_{1-\alpha}^2$ value. Testing $H_{0,joint}$ can be carried out by checking whether the null-hypothesized values of $(\beta_1, \beta_2)$ lie outside the ellipsoidal joint confidence region, in which case one rejects $H_{0,joint}$. If we want to make individual decisions based on this $H_{0,joint}$-derived ellipsoidal confidence region, we essentially construct two individual confidence intervals from the joint confidence region. One general way of carrying out this projection is to choose the bounds of the individual confidence intervals as the maximum value the joint confidence region attains in the associated dimension, which is represented by the light-gray lines enveloping the ellipsoid in Figure 2[4].

[4]The Scheffé (1953) method, based on inferring from an elliptical joint region onto individual hypotheses, is not optimal

The problem resulting from these individual decisions based on $H_{0,joint}$ is the following. The ellipsoid itself has good coverage properties and is powerful with respect to $H_{0,joint} : \beta_1 = \beta_2 = 0$. However, the projected individual $H_{0,joint}$-derived confidence intervals $CI_{\beta_1}^{joint}$ and $CI_{\beta_2}^{joint}$ are too large, in the sense that they have too little power against the individual hypotheses. Additionally, it is not clear whether using $CI_{\beta_1}^{joint}$ and $CI_{\beta_2}^{joint}$ controls the familywise error rate FWE at level $\alpha$.

# 4 False Discoveries in Estimating Multiple Treatment Effects

I now turn to reexamining the empirical labor market model of Lalive et al. (2005) with respect to false discoveries. For reasons of brevity, I describe their model only shortly in this section. Their competing risks duration model is outlined in more detail in Appendix A.

Lalive et al. (2005) analyze the treatment effects of benefit sanctions and corresponding warnings on the duration of unemployment. Having controlled for observable characteristics $x$ of an individual, let $\delta_1$ denote the treatment effect of a warning, and let $\delta_2$ denote the treatment effect of a benefit sanction on the duration of unemployment.

Common sense suggests that one tries harder to find a job when one is faced with less unemployment benefits. A warning of an unemployment benefit sanction may even be enough to induce the unemployed to look harder for jobs. If the benefit sanction is even enforced, there is a strong disutility from being unemployed, thus even the most recalcitrant unemployed have a strong incentive to find a job. This is formalized in the theoretical economic model of Boone and van Ours (2006). Lalive et al. (2005) quantify the two treatment effects of a warning, $\hat{\delta}_1$, and of a benefit sanction, $\hat{\delta}_2$, on the duration of unemployment. They use a large and reliable data set with 10,404 observations from Swiss labor market authorities to estimate the model.

In what follows, I look at cases in which the danger of making false discoveries is present. Lalive et al. (2005) do not take account of the danger of false discoveries, since they just report individual $t$-values. This corresponds to individually $t$-testing each coefficient. Given 10,404 observations, it seems reasonable to compare the observed $t$-statistics to quantiles from the standard normal distribution. The more coefficients that one tests individually, the higher is the danger of labeling nonexistent treatment effects as statistically significant, hence making false discoveries. Therefore, I will progressively increase the number of treatment coefficients under multiple statistical scrutiny within each of the following subchapters. In this step-wise manner, I will quantify the risk of making one or more false discoveries. Hence, we will see to which extent individual $t$-testing ignored the risk of labeling nonexistent treatment effects as statistically significant in the empirical context of Lalive et al. (2005)'s labor market model.

# 5 Empirical Results

A multiple test of the two treatment effects $\hat{\delta}_1$ (warning) and $\hat{\delta}_2$ (benefit sanction) as in (10) is a natural starting point

$$\{H_{0,1} : \delta_1 = 0, \quad H_{0,2} : \delta_2 = 0\}. \tag{10}$$

The difference between individual and multiple testing is not pronounced though, thus the results are not reported. This may be due to the low number of two null hypotheses or due to the correlation of

for FWE control either, as explained in detail in an oncoming paper from Wolf, Wunderli

the test statistics in the two null hypotheses, as I illustrated in Section 2.

## 5.1 Duration Dependence in Treatment Effects

### 5.1.1 Family of Null Hypotheses

If one accounts for flexible duration dependence of the treatment effects themselves, individual testing may find too many individually significant treatment effects. Table A.2 of Lalive et al. (2005) reports results of treatment effects taking account of duration dependence therein. Thus, the first multiple test scrutinizes the following family of null hypotheses while controlling the risk of false discoveries

$$\begin{aligned} \{H_{0,1} : \delta_{1,\text{'0-29 days'}} = 0, \quad H_{0,2} : \delta_{1,\text{'}\geq\text{30 days'}} = 0, \\ H_{0,3} : \delta_{2,\text{'0-59 days'}} = 0, \quad H_{0,4} : \delta_{2,\text{'}\geq\text{60 days'}} = 0\} \end{aligned} \tag{11}$$

$\delta_{1,\text{'0-29 days'}}$ denotes the treatment effect of a warning on the duration of unemployment during the first 29 days after the warning. 30 days or more after the warning was communicated to the unemployed, $\delta_{1,\text{'}\geq\text{30 days'}}$ denotes the incremental effect to $\delta_{1,\text{'0-29 days'}}$. Hence, $\delta_{1,\text{'}\geq\text{30 days'}} < 0$ does not mean that the effect of the warning after 30 days was to increase the duration of unemployment. It means that the overall effect of a warning after 30 days is $\delta_{1,\text{'0-29 days'}} + \delta_{1,\text{'}\geq\text{30 days'}}$, which is thus smaller than the treatment effect of a warning during the first 29 days $\delta_{1,\text{'0-29 days'}}$. Therefore, the family of alternative hypotheses to (11) needs to be two-sided as follows

$$\begin{aligned} \{H_{A,1} : \delta_{1,\text{'0-29 days'}} \neq 0, \quad H_{A,2} : \delta_{1,\text{'}\geq\text{30 days'}} \neq 0, \\ H_{A,3} : \delta_{2,\text{'0-59 days'}} \neq 0, \quad H_{A,4} : \delta_{2,\text{'}\geq\text{60 days'}} \neq 0\}. \end{aligned} \tag{12}$$

### 5.1.2 Testing Results

Table 3 summarizes the results from individual and multiple testing of (11) with two-sided alternative hypotheses. Each coefficient Coeff. in the first column is followed by its estimate $\widehat{\text{Coeff.}}$ and its $t$-statistic $|t|$. The columns labeled as individual tests list the critical values $c_{\cdot}$, the $p$-values $p^{\cdot}$, and if the individual test rejected at the 5% level, denoted as rej$^{\cdot}$. First under the normal distribution as in naive $CI^{N(0,1)}_{\beta_s}$ in (7), denoted as $c^{\mathcal{N}}_{0.975}$, $p^{\mathcal{N}}$, rej$^{\mathcal{N}}$. Second under bootstrapping as in naive $CI^{boot}_{\beta_s}$ in (8), denoted as $c^{boot}_{|.|,0.95}$, $p^{boot}$, rej$^{boot}$. Details of the one- and two-sided bootstrapping methodology are in Appendix B. Note that all tests are at the 5% level[5].

The columns labeled as 'multiple test' contain the multiple critical value $c^{MTest}_{|.|,0.95}$ such that the familywise error rate FWE defined in (2) is controlled at the 5% level. This means that we control the risk of labeling one or more nonexistent treatment effects as statistically significant (making one or more false discoveries) at the 5% level. The last column rej$^{MTest}$ indicates which individual null hypotheses can be rejected using the multiple test that controls the risk of making false discoveries.

First, observe that the first three coefficients are statistically significantly different from zero under the $\mathcal{N}$-distribution assumption, under bootstrapping, and under multiple testing. Thus, despite the fact that Lalive et al. (2005) just $t$-tested individually, they seem not having made false discoveries in the sense of having labeled nonexistent treatment effects as significant[6].

[5] $c^{\mathcal{N}}_{0.975} = c^{\mathcal{N}}_{|.|,0.95}$ corresponds to the $c^{boot}_{|.|,0.95}$ notation in Romano and Wolf (2005)

[6] We cannot match their reported results exactly, because we do not have the set of regressors $k$ denoting public employment service dummies (PES), see Appendix A.

| | | | individual tests at 5% level | | | | | | multiple test 5% level | |
|---|---|---|---|---|---|---|---|---|---|---|
| Coeff. | $\widehat{\text{Coeff.}}$ | $\lvert t\rvert$ | $c^{\mathcal{N}}_{0.975}$ | $p^{\mathcal{N}}$ | rej$^{\mathcal{N}}$ | $c^{boot}_{\lvert.\rvert,0.95}$ | $p^{boot}$ | rej$^{boot}$ | $c^{MTest}_{\lvert.\rvert,0.95}$ | rej$^{MTest}$ |
| $\delta_{1,\text{'0-29 days'}}$ | 0.4103 | 5.51 | 1.96 | 0.000 | yes | 2.00 | 0.000 | yes | 2.40 | yes |
| $\delta_{1,\text{'}\geq\text{30 days'}}$ | −0.2522 | 3.25 | 1.96 | 0.000 | yes | 2.05 | 0.012 | yes | 2.40 | yes |
| $\delta_{2,\text{'0-59 days'}}$ | 0.2925 | 2.47 | 1.96 | 0.007 | yes | 2.17 | 0.032 | yes | 2.40 | yes |
| $\delta_{2,\text{'}\geq\text{60 days'}}$ | 0.0437 | 0.42 | 1.96 | 0.338 | no | 1.97 | 0.701 | no | 2.40 | no |

Table 3: Results from testing four treatment effects under two-sided alternatives

Second, note that the $\mathcal{N}$-distribution assumption seems quite accurate, given that the bootstrap critical values $c^{boot}_{|.|,0.95}$ and the $\mathcal{N}$ derived $c^{\mathcal{N}}_{0.975}$ are quite close to each other. This does not hold for the associated $p$-values $p^{\mathcal{N}}$ and $p^{boot}$, though. Also note that the individual critical values $c^{boot}_{|.|,0.95}$ differ from each other, while the critical value resulting from multiple testing is one and the same by construction for all coefficients.

It is possible to control more liberal multiple error types I, such as the 2-FWE being the probability of making two or more false discoveries; the interested reader finds these testing results in Appendix C.

### 5.1.3 Quantifying the Risk of Making False Discoveries by Individual Testing

The bootstrapping of the model produced slightly different individual critical values for the $t$-tests than by using the $\mathcal{N}$-assumption. That the difference between assuming $\mathcal{N}$ and bootstrapping is not pronounced is not surprising: there are $10,404$ data points. This conclusion was drawn in Table 3 by comparing the individual $\mathcal{N}$-assumed critical values 1.96 with the individual bootstrap critical values $c^{boot}_{|.|,0.95}$ that range from 1.97 to 2.17.

How large is the risk of making one or more false discoveries[7] by individual testing? It turns out that one needs to be ready to run the risk of making one or more false discoveries at around 20% to justify the $\mathcal{N}$-assumed critical value of 1.96. Thus, by individually testing at the 5% significance level under $\mathcal{N}$, the implicit risk of making one or more false discoveries is around 20%. This is clearly too large a multiple error type one to justify it with conventional significance levels. Hence, conventional significance levels at the individual testing level do not translate into putting conventional thresholds on multiple error types one that take account of the danger of false discoveries.

**Note:** In principle, this high risk of making false discoveries can be attributed to the following two sources

1. Individual bootstrap testing versus multiple testing
2. Assumed normality versus data-driven approximation of the data generating distribution by bootstrapping

Given that there are $10,404$ i.i.d. data points, the sampling distribution should be close to the normal distribution as a central limit theorem suggests. Thus, the main source of this high risk of making false discoveries should be the naive individual testing instead of the multiple test. Quantifying the impor-

[7]That is, of falsely labeling one or more treatment effects that do not exist as statistically significant

tance of these two sources exactly requires a very computing-intensive two-stage bootstrap simulation, as shortly described in the note at the end of Appendix B.

## 5.2 Qualification Dependence in Treatment Effects

### 5.2.1 Family of Null Hypotheses

A natural question to ask is whether the effect of warnings or benefit sanctions on the duration of unemployment depends on the qualification of an unemployed person. One may also expect a systematic pattern in treatment effects with respect to gender, age groups, or with respect to other explanatory variables. I choose to differentiate the two treatment effects $\delta_1$, $\delta_2$ with respect to three levels of qualification *quali*1, *quali*2, *quali*3 here, resulting in six treatment effect coefficients. The family of null hypotheses containing the six treatment effects is

$$\begin{aligned}\{H_{0,1} : \delta_{1,\text{'quali1'}} = 0, \quad H_{0,2} : \delta_{1,\text{'quali2'}} = 0, \quad H_{0,3} : \delta_{1,\text{'quali3'}} = 0,\\ H_{0,4} : \delta_{2,\text{'quali1'}} = 0, \quad H_{0,5} : \delta_{2,\text{'quali2'}} = 0, \quad H_{0,6} : \delta_{2,\text{'quali3'}} = 0\}\end{aligned} \tag{13}$$

Note that there are no incremental effects here as in the case of duration dependence in treatment effects. This means that it does not make economic sense if any of these qualification differentiated treatment effects is negative. Hence, the family of alternative null hypotheses is one-sided

$$\begin{aligned}\{H_{A,1} : \delta_{1,\text{'quali1'}} > 0, \quad H_{A,2} : \delta_{1,\text{'quali2'}} > 0, \quad H_{A,3} : \delta_{1,\text{'quali3'}} > 0,\\ H_{A,4} : \delta_{2,\text{'quali1'}} > 0, \quad H_{A,5} : \delta_{2,\text{'quali2'}} > 0, \quad H_{A,6} : \delta_{2,\text{'quali3'}} > 0\}.\end{aligned} \tag{14}$$

### 5.2.2 Testing Results

The results from these qualification differentiated treatment effects (13) are listed in Table 4.

Each coefficient Coeff. in the first column is followed by its estimate $\widehat{\text{Coeff.}}$, and its $t$-statistic labeled as $t$. The columns labeled as individual tests list the critical values $c^{\cdot}$, $p$-values $p^{\cdot}$, and rejection at the 5% level $\text{rej}^{\cdot}$ of the individual tests. First under the normal distribution $c^{\mathcal{N}}_{0.95}$, $p^{\mathcal{N}}$, $\text{rej}^{\mathcal{N}}$, and second under bootstrapping $c^{boot}_{0.95}$, $p^{boot}$, $\text{rej}^{boot}$.

The columns labeled as 'multiple test' contain the multiple critical value $c^{MTest}_{0.95}$ such that the multiple error type one 'probability of falsely rejecting one or more null hypotheses' is controlled at the 5% level. This means that I control the risk of labeling one or more nonexistent treatment effects as statistically significant (one or more false discoveries) at the 5% level. The last column $\text{rej}^{MTest}$ indicates which individual null hypotheses can be rejected using the multiple test instead of the individual tests.

$\delta_{1,\text{'quali3'}}$, $\delta_{2,\text{'quali3'}}$, and $\delta_{1,\text{'quali2'}}$ are found to be significantly larger than zero under individual testing with an $\mathcal{N}$ assumption. Under individual testing using bootstrapping instead of the $\mathcal{N}$ assumption, one also labels three treatment effect as statistically significant. However, these three individually significant treatment effects from individual testing are probably false discoveries, as multiple testing at the 5% significance level indicates. Under multiple testing, none of the six treatment effects are found to be statistically significantly greater than zero[8]. Thus, individual tests seem to falsely label these treatment effects as statistically significant at the 5% level, while multiple testing indicates that these are false discoveries at the 5% level.

[8]Nonetheless, by controlling the FWE at the 10% level rather than at the 5% level, the three treatment effects that are significant under individual bootstrap testing remain statistically significant under multiple testing.

| | | | individual tests at 5% level | | | | | | multiple test 5% level | |
|---|---|---|---|---|---|---|---|---|---|---|
| Coeff. | $\widehat{\text{Coeff.}}$ | $t$ | $c_{0.95}^{\mathcal{N}}$ | $p^{\mathcal{N}}$ | rej$^{\mathcal{N}}$ | $c_{0.95}^{boot}$ | $p^{boot}$ | rej$^{boot}$ | $c_{0.95}^{MTest}$ | rej$^{MTest}$ |
| $\delta_{1,\text{'quali3'}}$ | 0.3282 | 4.39 | 1.64 | 0.000 | yes | 0.783 | 0.000 | yes | 4.43 | no |
| $\delta_{2,\text{'quali3'}}$ | 0.1805 | 1.70 | 1.64 | 0.044 | yes | 1.233 | 0.019 | no | 4.43 | no |
| $\delta_{1,\text{'quali2'}}$ | 0.3795 | 4.28 | 1.64 | 0.000 | yes | 0.687 | 0.002 | yes | 4.43 | no |
| $\delta_{2,\text{'quali2'}}$ | 0.1855 | 1.40 | 1.64 | 0.081 | no | 1.344 | 0.044 | yes | 4.43 | no |
| $\delta_{1,\text{'quali1'}}$ | 0.0409 | 0.80 | 1.64 | 0.211 | no | 4.371 | 0.953 | no | 4.43 | no |
| $\delta_{2,\text{'quali1'}}$ | 0.0020 | 0.02 | 1.64 | 0.490 | no | 2.954 | 0.910 | no | 4.43 | no |

Table 4: Results from testing six treatment effects under one-sided alternatives

#### 5.2.3 Quantifying the Risk of Making False Discoveries by Individual Testing

Here again, the question is of what magnitude the risk of making false discoveries is by testing only individually. Note that the multiple critical value 4.43 puts a 5% threshold on the familywise error rate[9]. The individual critical value under $\mathcal{N}$ at 1.64 is very far off the multiple critical value 4.43.

It is thus not surprising that the risk of making one or more false discoveries is around an unacceptably high 90% if one tests individually under the $\mathcal{N}$ distribution. The same note as in Subsection 5.1.3 concerning the two sources of the risk of making false discoveries applies here.

## 6 Conclusions

Is is important to test multiple effects in a multiple testing manner to guard against the danger of making false discoveries. If we test coefficients only individually by looking at their $t$-statistics, we run the danger of so-called false discoveries. That is, we run the danger of labeling treatment effects as statistically significant that do not exist. The simulation study illustrated that the higher the number of coefficients is that one looks at jointly, the higher is the risk of making such false discoveries. Furthermore, the lower the correlation is between the random variables associated with the null hypotheses, the higher is the risk of making one or more false discovery.

To this end, the well-known $F$-test or the more general Wald, Lagrange Multiplier, or Likelihood Ratio test have one major shortcoming. Any of these joint tests can essentially test just one *joint* null hypothesis against its *joint* alternative hypothesis. Thus, any of these tests can only tell us in special cases which individual coefficients contained in the joint null hypothesis are significant from a joint point of view; multiple testing methods generally allow for individual rejections.

The study from Lalive et al. (2005) that I reexamine uses a data set of 10,404 independent observations. Individual testing at the 5% significance level under the normal distribution translates into high risks of making false discoveries. Namely, the probabilities of falsely labeling one or more nonexistent treatment effects as statistically significant is around 20%[10] or even around 90%[11].

Multiple testing methods allow putting multiple treatment effects under joint statistical scrutiny, while controlling the risk of making false discoveries. Lalive et al. (2005) do not seem to have reported

[9] I.e., making one or more false discoveries

[10] Four treatment effect coefficients, hence four null hypotheses

[11] Six treatment effect coefficients, hence six null hypotheses

false discoveries, despite the fact that they just tested their treatment effects individually.

However, by differentiating treatment effects of benefit sanctions on the basis of qualifications[11], I provide evidence that individual $t$-tests probably make three false discoveries. That is, individual testing labels three treatment effects out of six as statistically significant. By putting a 5% threshold on the risk of making one or more false discoveries, these three individually significant treatment effects are indicated as false discoveries by multiple testing methods from Romano and Wolf (2005).

Unfortunately, most applied work does not take the risk of false discoveries into account, since only individual $t$-statistics are reported, or a test of one joint null hypothesis at the most. Among others, this paper and Heckman et al. (2010) highlight the need to control the risk of making false discoveries. From a meta point of view, the problem of false discoveries is even aggravated. If scientists only report results that work out of an actually much larger pool of candidate results they have tried, which Heckman et al. (2010) refer to as cherry picking, the danger of selectively reporting convenient results that may in fact be false discoveries undermines scientific credibility.

## References


Boone, J. and van Ours, J. C. (2006). Modelling financial incentives to get unemployed back to work. *Journal of Institutional and Theoretical Economics*, 162(2):227–252.

Heckman, J., Moon, S. H., Pinto, R., Savelyev, P., and Yavitz, A. (2010). Analyzing social experiments as implemented: A reexamination of the evidence from the highscope perry preschool program. *Quantitative Economics*, 1(1):1–46.

Heckman, J. J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, 52:271–320.

Lalive, R., van Ours, J. C., and Zweimueller, J. (2005). The effect of benefit sanctions on the duration of unemployment. *Journal of the European Economic Association*, 3(6):1386–1417.

Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

Scheffé, H. (1953). A method for judging all constrasts in the analysis of variance. *Biometrika*, 40:87–104.

# A The Model that I Reexamine with Respect to False Discoveries

Let the random variable $T_u$ denote the duration spent in unemployment. $T_{s1}$ denotes the duration from entry to unemployment until a person gets a warning. Let $T_{s2}$ denote the duration from a warning until the 100% benefit sanction is enforced, which is the case under Swiss law. The corresponding rates (15), (16), (17) of $T_u$, $T_{s1}$, $T_{s2}$ are parameterized with observables only or with unobservables to allow for unobserved heterogeneity over individuals. In the first model without unobservables, Lalive et al. (2005) assume that the three rates of $T_u$, $T_{s1}$, $T_{s2}$ are explained perfectly well by a set of observable variables. To make the model more realistic, they add unobserved heterogeneity terms $u$, $v_1$, $v_2$ within the rates $\theta_u$, $\theta_{s1}$, $\theta_{s2}$ as follows

$$\theta_u(t \mid x, D_1, D_2, u) = \lambda_u(t) \exp\{x'\beta_u + \delta_1 D_1 + \delta_2 D_2 + u\}, \tag{15}$$

$$\theta_{s_1}(t \mid x, v_1) = \lambda_{s_1}(t) \exp\{x'\beta_{s_1} + v_1\}, \tag{16}$$

$$\theta_{s_2}(t \mid x, v_2) = \lambda_{s_2}(t) \exp\{x'\beta_{s_2} + v_2\}. \tag{17}$$

which corresponds to equations (11) and (12) in Lalive et al. (2005).

(15) is the rate at which individuals drop out of employment. The higher $\theta_u$ is, the more likely is the favorable case that the individual drops out of unemployment. (16) is the rate of getting a warning. (17) is the rate of getting an unemployment benefit sanction. $x$ denotes individual characteristics. Lalive et al. (2005) have an additional set of regressors $k$, which are public employment dummy variables. We could not get the data of these dummy variables $k$, thus our results do not perfectly match theirs. $D_1 = I(t_{s1} < t_u)$ is a dummy variable indicating if there was a warning for the individual. $D_2 = I(t_{s2} < t_u)$ denotes the dummy variable indicating if a sanction was enforced for the individual. The coefficients $\delta_1$ and $\delta_2$ are the two so-called ex-post treatment effects of benefit sanctions. Due to our missing public employment service dummy variables, we cannot estimate ex-ante treatment effects as in Lalive et al. (2005), which is a key innovation of their paper.

The $\lambda_\bullet(t)$ are coefficients, modeling flexible duration dependence with a step function. Generally speaking, the longer a person is unemployed, the less likely finding a job becomes, thus $\lambda_u(t)$ models negative duration dependence.

$$\lambda_u(t) = \exp\{\sum_{k=0}^{4} \lambda_{u,k} I_k(t)\}, \quad \lambda_{s1}(t) = \exp\{\sum_{k=0}^{4} \lambda_{s1,k} I_k(t)\}, \quad \lambda_{s2}(t) = \exp\{\sum_{k=0}^{4} \lambda_{s2,k} I_k(t)\} \tag{18}$$

The indicator functions $I_{\bullet,k}(t)$ are set up for the following time intervals, respectively: 0 to 3 months, 3 to 6 months, 6 to 9 months, 9 to 12 months, 12 and more months. Each $\lambda_{\bullet,0}$ is set to zero because a constant term is also estimated.

The model comprises a competing risks specification. That is, while a person is unemployed, he runs the competing risks of getting a benefit sanction warning or finding a job. Once the person has got a warning, he runs the competing risks of a benefit sanction or finding a job. Once the person incurred a benefit sanction, he is left with the single risk of finding a job.

There are four treatment effect coefficients in Section 5.1 of this paper, which enter the three rates

as follows

$$\theta_u(t \mid x, D_1, D_2, u) = \lambda_u(t) \exp\{x'\beta_u + \delta_{1,\text{0-29d}} D_{1,\text{0-29d}} + \delta_{1,\geq 30\text{d}} D_{1,\geq 30\text{d}} + \delta_{2,\text{0-59d}} D_{2,\text{0-59d}} + \delta_{2,\geq 60\text{d}} D_{2,\geq 60\text{d}} + u\}, \tag{19}$$

$$\theta_{s_1}(t \mid x, v_1) = \lambda_{s_1}(t) \exp\{x'\beta_{s_1} + v_1\}, \tag{20}$$

$$\theta_{s_2}(t \mid x, v_2) = \lambda_{s_2}(t) \exp\{x'\beta_{s_2} + v_2\}. \tag{21}$$

In Section 5.2 of this paper, I differentiated treatment effects based on three levels of qualifications 'quali1', 'quali2', 'quali3', which are abbreviated as $q1$, $q2$, $q3$ here, respectively. Thus, there are three rates $\theta_{u,q1}$, $\theta_{u,q2}$, $\theta_{u,q3}$ instead of just one rate $\theta_u$ as before. The resulting five rates are

$$\theta_{u,q1}(t \mid x, D_1, D_2, u_{q1}) = \lambda_{u,q1}(t) \exp\{x'\beta_{u,q1} + \delta_{1,q1} D_1 + \delta_{2,q1} D_2 + u_{q1}\}, \tag{22}$$

$$\theta_{u,q2}(t \mid x, D_1, D_2, u_{q2}) = \lambda_{u,q2}(t) \exp\{x'\beta_{u,q2} + \delta_{1,q2} D_1 + \delta_{2,q2} D_2 + u_{q2}\}, \tag{23}$$

$$\theta_{u,q3}(t \mid x, D_1, D_2, u_{q3}) = \lambda_{u,q3}(t) \exp\{x'\beta_{u,q3} + \delta_{1,q3} D_1 + \delta_{2,q3} D_2 + u_{q3}\}, \tag{24}$$

$$\theta_{s_1}(t \mid x, v_1) = \lambda_{s_1}(t) \exp\{x'\beta_{s_1} + v_1\}, \tag{25}$$

$$\theta_{s_2}(t \mid x, v_2) = \lambda_{s_2}(t) \exp\{x'\beta_{s_2} + v_2\}. \tag{26}$$

Lalive et al. (2005) estimate the model by maximizing the resulting closed-form log-likelihood function. The Heckman-Singer mass point approach is used to estimate the model with unobservables by maximum likelihood, as in Heckman and Singer (1984). I used TSP (Time Series Package) to estimate the model, as Lalive et al. (2005) did. The ML solver of TSP can make use of analytic derivatives, which is a neat feature for the closed-form densities of the model. These authors did not bootstrap the model, however, they rely on asymptotic normality to judge the significance of the estimated treatment effects.

# B Implementation

The short code of the simulation study is available from the author on request; I do not elaborate on it here. Nonetheless, I elaborate on the implementation of the empirical part of this paper, which consisted of the following five steps

(a) Replicate Lalive et al. (2005)'s results

(b) Implement four and six treatment effects model based on replicating code

(c) Bootstrap the four and six treatment effects model

(d) Do individual and multiple testing of four or six treatment effects based on bootstrap results

(e) Quantify the risk of making one of more false discoveries by individual instead of multiple testing

## B.1 Replicate their results

I got the original TSP code (Time Series Package) for the basic two treatment effects model in Lalive et al. (2005) from Raphael Lalive, to which I could compare my implementation. He also provided me with the original data, except for the public employment dummies, which he was not allowed to pass on to me due to data protection issues.

## B.2 Implement Four and Six Treatment Effects Model Based on Replicating Code

Rafael Lalive also helped me set up the code for the four treatment effects model that is contained in Lalive et al. (2005). Based on these two codes, I coded the six treatment effects model. Details of the six treatment effects model can be found in Appendix A.

## B.3 Bootstrap the Four and Six Treatment Effects Model

Bootstrapping my two models was fairly easy, given that one can use the conventional i.i.d. bootstrap by drawing single data points with replacement from the original data.

Specifically, let $\hat{\theta}$ denote the ML coefficients of one of my two models using the original data

$$\hat{\theta} = \arg\max_{\theta} L(\theta; data), \tag{27}$$

where $data$ denotes the $[10404 \times p]$ matrix containing the original data and $L(\cdot;\cdot)$ denotes the (log) likelihood. Given no dependence between individuals but possible contemporaneous dependence between variables, one can generate an artificial bootstrap data set $data_1^*$ by drawing single data rows with replacement $10,404$ times from the sequence of data rows $1, 2, \ldots, 10404$. Note that by falsely i.i.d bootstrapping each variable, that is column, separately, the possible contemporaneous dependence of the variables is destroyed. Thus, the first bootstrap data set of size $[10404 \times p]$ may be

$$\text{First bootstrap data set } data_1^* : \underbrace{\text{data row}_{235}, \text{data row}_{52}, \text{data row}_{9874}, \ldots, \text{data row}_{52}}_{\text{10,404 data rows as in original data set}}$$

The second bootstrap data set of size $[10404 \times p]$ may look like

$$\text{Second bootstrap data set } data_2^* : \underbrace{\text{data row}_{189}, \text{data row}_{8532}, \text{data row}_{10203}, \ldots, \text{data row}_{9874}}_{\text{10,404 data rows as in original data set}}$$

In this way, I generated $2,500$ bootstrap data sets $data_1^*$, $data_2^*$, ..., $data_{2500}^*$. By computing the ML coefficients on each of these 2,500 bootstrap data sets, I get 2,500 bootstrap ML coefficients

$$\begin{aligned} \hat{\theta}_1^* &= \arg\max_{\theta} L(\theta; data_1^*), \\ \hat{\theta}_2^* &= \arg\max_{\theta} L(\theta; data_2^*), \\ &\vdots \\ \hat{\theta}_{2500}^* &= \arg\max_{\theta} L(\theta; data_{2500}^*). \end{aligned} \tag{28}$$

## B.4 Do Individual and Multiple Testing of Four or Six Treatment Effects Based on Bootstrap Results

### B.4.1 Individual One- and Two-Sided Bootstrap Tests

To carry out an individual $t$-test of an individual coefficient $\beta \in \theta$ based on bootstrapping instead of an $\mathcal{N}$-assumption, one computes the $t$-value of the individual coefficient $\beta$ on each of these bootstrap data sets, resulting in 2,500 $t$-values $\hat{\beta}_1^*/\hat{\sigma}(\hat{\beta}_1^*)$, $\hat{\beta}_2^*/\hat{\sigma}(\hat{\beta}_2^*)$, ..., $\hat{\beta}_{2500}^*/\hat{\sigma}(\hat{\beta}_{2500}^*)$[12].

[12]The standard deviation $\hat{\sigma}(\hat{\beta}_m^*)$ on bootstrap data set $\text{data}_m^*$ was computed by the Eicker-White method, which is a combination of analytic second derivatives and the covariance of the analytic gradient. Asymptotically, these two ways of computing the standard deviation of an ML estimator obtain the same result, as stipulated in the so-called information matrix equality. On the computer, these two ways may yield different results, though. The Eicker-White estimator finds an optimal combination thereof. This corresponds to the `HCOV=W` option in TSP's `ML()` routine.

The bootstrap critical value at significance level $\alpha$ for a one-sided bootstrap test with alternative $H_A$ : $\beta > 0$ is just the $1-\alpha$ empirical quantile of the 2,500 $t$-values $\hat{\beta}_1^*/\hat{\sigma}(\hat{\beta}_1^*)$, $\hat{\beta}_2^*/\hat{\sigma}(\hat{\beta}_2^*)$, ..., $\hat{\beta}_{2500}^*/\hat{\sigma}(\hat{\beta}_{2500}^*)$.

The bootstrap critical value at significance level $\alpha$ for a two-sided bootstrap test with alternative $H_A : \beta \neq 0$ is the $1-\alpha$ empirical quantile of the 2,500 $t$-values in absolute value $\left|\hat{\beta}_1^*/\hat{\sigma}(\hat{\beta}_1^*)\right|$, $\left|\hat{\beta}_2^*/\hat{\sigma}(\hat{\beta}_2^*)\right|$, ..., $\left|\hat{\beta}_{2500}^*/\hat{\sigma}(\hat{\beta}_{2500}^*)\right|$.

If the observed $t$-value $\hat{\beta}/\hat{\sigma}(\hat{\beta})$ is larger than the $1-\alpha$ bootstrap critical value based on the 2,500 bootstrap data sets, the null hypothesis is rejected at significance level $\alpha$.

The one-sided bootstrap $p$-value for coefficient $\beta \in \theta$ is computed as

$$p^{boot} = \frac{\sum_{m=1}^{2500} \mathbf{1}\{\hat{\beta}_m^*/\hat{\sigma}^*(\hat{\beta}_m^*) > \hat{\beta}/\hat{\sigma}(\hat{\beta})\}}{2500+1}, \tag{29}$$

where **1** denotes the indicator function. In the two-sided case, the absolute values of the bootstrap and the original $t$-values must be used to compute the bootstrap $p$-value.

### B.4.2 Multiple Testing

The implementation of the multiple test is straight forward. On Michael Wolf's webpage, there is R and Matlab code available that carries out multiple testing. One can simply pass the vector of observed $t$-statistics $[\hat{\beta}/\hat{\sigma}(\hat{\beta}), \beta \in \theta]$ and the matrix of bootstrap $t$-statistics $[\hat{\beta}_1^*/\hat{\sigma}(\hat{\beta}_1^*),\ \hat{\beta}_2^*/\hat{\sigma}(\hat{\beta}_2^*),\ \ldots,\ \hat{\beta}_{2500}^*/\hat{\sigma}(\hat{\beta}_{2500}^*), \beta \in \theta]$ to the R or Matlab code. The bootstrap matrix $[\hat{\beta}_1^*/\hat{\sigma}(\hat{\beta}_1^*),\ \hat{\beta}_2^*/\hat{\sigma}(\hat{\beta}_2^*),\ \ldots,\ \hat{\beta}_{2500}^*/\hat{\sigma}(\hat{\beta}_{2500}^*), \beta \in \theta]$ consists of 2,500 rows, each row containing the estimated $t$-statistic for the $m^{th}$ bootstrap data set. As said before, for the two-sided alternative hypothesis case, element-wise absolute values must be provided.

Thus, for my four treatment effects model, I passed a $[4 \times 1]$ vector of observed $t$-statistics and a $[2500 \times 4]$ matrix of bootstrap $t$-statistics to the R function, since $[\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4] = \hat{\theta}$. For the six treatment effects model, I passed a $[6 \times 1]$ vector of observed $t$-statistics and a $[2500 \times 6]$ matrix of bootstrap $t$-statistics to the R function.

To control the familywise error rate at the 5% level, one must additionally provide the R or Matlab function with $k = 1$ and $\alpha = 0.05$. The R or Matlab function then returns the multiple critical value and the indices of the null hypotheses that could be rejected while controlling the familywise error rate at the $\alpha$% level.

## B.5 Quantify the Risk of Making One of More False Discoveries by Testing Individually Instead of Multiple Testing

Quantifying the risk of making one or more false discoveries by assuming $\mathcal{N}$ instead of relying on multiple testing is easy. The multiple testing method of Romano and Wolf (2005) based on bootstrapping essentially solves the following problem

MTest: Find one critical value $c_{1-\alpha}^{MTest}$ such that familywise error rate FWE $\leq \alpha$ asymptotically.

Quantifying the probability of making one or more false discoveries $\widehat{\text{FWE}}$ in assuming $\mathcal{N}$ essentially solves the related problem

Find $\widehat{\text{FWE}}$: Find $\hat{\alpha}$ such that the $\mathcal{N}$-assumed critical value $c_{1-\alpha}^{\mathcal{N}\text{ass.}}$ ensures FWE $\leq \hat{\alpha}$ asymptotically.

This can be done by a grid search over the parameter $\alpha$ that is passed on to the multiple testing routine. For example, in the two-sided testing, the $\mathcal{N}$ assumption results in the critical value 1.96 at the 5% level. I quantify the committed multiple error type one $\widehat{\text{FWE}}$ in using 1.96 instead of the multiple critical value as follows. Increase $\hat{\alpha}$ passed to the multiple testing routine until a multiple critical value of 1.96 is returned. The $\hat{\alpha}$ that satisfies this criterion up to two decimal points is the approximate risk of making one or more false discoveries $\widehat{\text{FWE}}$ in assuming $\mathcal{N}$ instead of relying on multiple testing.

**Note: Two sources of error in individually testing under normal assumption** It would be very interesting to know exactly whether the high risk of making one or more false discoveries by assuming $\mathcal{N}$ is mostly due to individual bootstrap testing instead of multiple bootstrap testing. Or whether it is mainly due to assuming $\mathcal{N}$ instead of bootstrapping the data generating distribution. I expect the error stemming from assuming $\mathcal{N}$ instead of bootstrapping to be small, as a central limit theorem suggests for $10,404$ data points. Nonetheless, to answer this question exactly, the easy approximation scheme as in step (e) does not work for this case, since there is not a single individual bootstrap critical value covering all null hypotheses. For example, the individual two-sided bootstrap critical values $c^{boot}_{|.|,0.95}$ range from 1.97 to 2.17 for the four treatment effects model.

Hence, one needs a two-stage bootstrap analysis to answer this question, which is very computing intensive for this nonlinear ML problem. Specifically, one needs not only carry out an ML routine on each of the 2,500 bootstrap data sets, as was the case to carry out multiple testing. One even needs to conduct a so-called second-stage bootstrap analysis of 500 repetitions, say, on each of the 2,500 first-stage bootstrap samples, to answer this question. Hence, the ML problem needs to be solved $2,500 \times 501$ times, which takes a long time of parallel computing.

# C Control of More Liberal Multiple Error Types

**Four treatment effects: Duration Dependence in Treatment Effects** Note that I control the 'probability of falsely rejecting one or more null hypotheses' at the 5% level.

What happens if we get more liberal with respect to false discoveries, thus merely want to control the 'probability of falsely rejecting *two or more* null hypotheses'? What if we even considered the very liberal 'probability of falsely rejecting *four or more* null hypotheses'?

What essentially happens is that the more liberal the multiple error type one gets in the aforementioned sense, the lower the critical value gets that the multiple testing method returns. Controlling the 'probability of falsely rejecting *two or more* null hypotheses' at the 5% level is achieved by a multiple critical value of 1.36. The two multiple critical values 1.00 and 0.80 control the multiple error types I '*three or more ...*' and '*four or more ...* ' at the 5% level, respectively. Thus, the higher one sets the number of false discoveries in the risk threshold, the lower the critical value gets, thus the more treatment effects are labeled as statistically significant. By merely individually testing, one knows that the risk of false discoveries is present, but one cannot put a risk threshold on it. Unfortunately, this seems to be the modus operandi in most applied work.

One can avoid choosing the $k$ in controlling the 'probability of falsely rejecting $k$ *or more* null hypotheses' by considering *relative* multiple error types I, such as the False Discovery Proportion (FDP). Control of the FDP is achieved by increasing $k$ within 'probability of falsely rejecting $k$ or more null hypotheses' until a criterion is met; see Romano and Wolf (2005) for details.

**Six treatment effects: Qualification Dependence in Treatment Effects** The more liberal probability of falsely declaring two or more false discoveries is still at 33.3% by the individual critical value 1.64 resulting from assuming $\mathcal{N}$. Recall that the probability of falsely rejecting one or more null hypotheses was around 90%.

The $\mathcal{N}$ derived individual critical value 1.64 puts a 1.9% threshold on the very relaxed multiple error type one 'probability of falsely declaring three or more treatment effects as statistically significant'. Thus, it is perfectly possible that by individually testing, one does control the risk of making a number of false discoveries at conventional significance levels by coincidence. But again, multiple testing methods let us quantify and put a threshold on the risk of making false discoveries.